INTERLEAVING AND ESTIMATION OF LOST VECTORS FOR ROBUST SPEECH RECOGNITION IN BURST-LIKE PACKET LOSS

A.B. James, B.P. Milner

School of Computing Sciences, University of East Anglia, Norwich, U.K.
email: {a.james, b.milner}@uea.ac.uk web: www.cmp.uea.ac.uk

ABSTRACT

Analysis into the effect of packet loss on speech recognition performance shows that both the burst length and the overall proportion of packets lost contribute to a deterioration in accuracy. To combat this burst-like packet loss several methods are compared for estimating the value of missing feature vectors. Three forms of interleaver are then compared which distribute long duration bursts of packet loss into a series of smaller bursts in the feature vector stream. Experimental results are presented on a range of channel conditions and demonstrate that substantial accuracy gains can be achieved using estimation techniques provided burst lengths are short. For longer burst lengths interleaving is necessary to maintain performance. For example at a packet loss rate of 50% and average burst length 20 packets (which represents 40 feature vectors or 400ms) performance is increased from 49.6% with no compensation to 86% with interleaving and cubic interpolation.

1. INTRODUCTION

The move towards mobile and handheld devices for speech communication has lead to distributed speech recognition (DSR) systems being developed. The Aurora DSR standard proposed by the European Telecommunication Standards Institute (ETSI) offers good robustness to noise by replacing the low bit-rate speech codec on the terminal device with the static MFCC feature extraction component of the speech recogniser [1]. Figure 1 shows an overview of a typical DSR system along with the proposals outlined in this work.

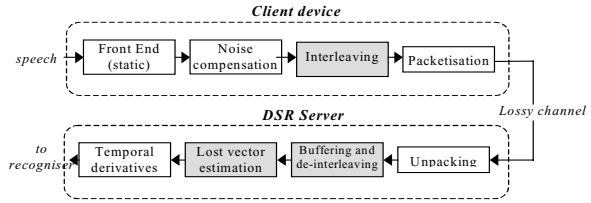

Figure 1: Architecture of the proposed DSR system.

The networks across which DSR systems transmit packetised speech data often do not guarantee reliable delivery. When packet loss occurs, or too many bits are corrupted so that bit level forward error correction cannot correct the frame, then portions of the feature vector stream become lost. Early work on packet loss compensation for DSR considered splicing the feature vector stream together in loss periods [2] or duplicating correctly received vectors to compensate for lost vectors [1,3]. Alternative schemes have used interpolation to estimate lost packets [4] or have added error correction bits to protect the speech data [5]. These schemes have varying degrees of success and work reasonably well for short duration bursts of loss but degrade as burst lengths increase.

The conditions that cause packet loss on both mobile and IP networks often have sufficient duration to cause bursts of loss to occur. Therefore, to characterise a channel in terms of its packet loss, two metrics need to be considered; the proportion of packets lost, α, and the average burst length, β. Figure 2 shows how these two characteristics affect speech recognition accuracy for packet loss rates from 10% to 50% and average burst lengths from 1 to 20 vectors – see section 4 for experimental details. No packet loss compensation is employed in figure 2a with the result that accuracy is largely governed by the packet loss rate, α, whilst the average burst length, β, has far less effect. It is interesting to observe that as burst length increases, the accuracy converges to:

baseline accuracy × (1 – proportion of vectors lost)

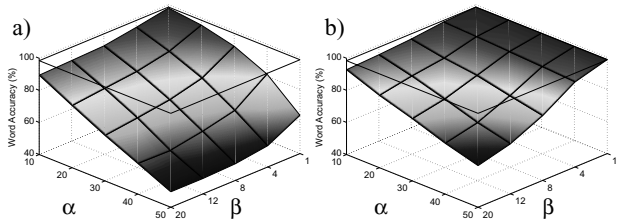

Figure 2: Word accuracy against varying channel condition with: a) no compensation , b) interpolation.

Increased accuracy can be achieved by estimating the missing vectors prior to recognition. This is shown in figure 2b which applies linear interpolation to estimate the value of lost vectors. Now the overall loss rate, α, has less effect on accuracy than the average burst length, β. This is because interpolation is able to correct short duration bursts of loss but is less effective at estimating missing vectors which occur in longer bursts. This indicates that when estimating lost vectors it is not the proportion of vectors lost that is significant, but rather the average burst length. Indeed, baseline accuracy of 98.6% can be maintained even at a loss rate of 50% providing the average burst length is short. Thus it is more important to reduce the average burst length rather than reduce the overall packet loss rate through channel coding schemes. This work considers the combination of methods for estimating missing vectors and interleaving to reduce average burst lengths for distributed speech recognition.

2. ESTIMATION OF LOST VECTORS

The estimation methods considered in this section use vectors that were correctly received before and after the burst of packet

loss to determine the value of lost vectors. The number of vectors used before and after the burst are labelled N_{before} and N_{after}. Two forms of estimation are considered; repetition and interpolation. Repetition simply replaces lost vectors with copies of vectors immediately surrounding the loss. Interpolation fits a mathematically defined curve onto correctly received vectors surrounding the loss from which lost vectors can be estimated.

2.1 Repetition

In its most elementary form, repetition simply replaces a missing vector with a copy of the most recent correctly received vector, $\mathbf{x}_{\text{before}}$. Therefore the estimate of the n^{th} vector of the burst is given,

$$\hat{\mathbf{x}}_n = \mathbf{x}_{\text{before}} \qquad 1 \le n \le \beta \qquad (1)$$

As only one previous vector is required, $N_{before}=1$ and $N_{after}=0$. An improvement can be made by replacing each missing vector with a copy of the vector received either before or after the burst, depending on which it is closest,

$$\hat{\mathbf{x}}_n = \begin{cases} \mathbf{x}_{\text{before}} & n < \beta/2 \\ \mathbf{x}_{\text{after}} & n \ge \beta/2 \end{cases} \qquad 1 \le n \le \beta \qquad (2)$$

where $\mathbf{x}_{\text{before}}$ and $\mathbf{x}_{\text{after}}$ are the vectors received before and after the loss respectively and β is the burst length. This method requires $N_{before}=1$ and $N_{after}=1$. This technique is known as *nearest neighbour repetition* and is the principle method of vector loss compensation specified in the ETSI Aurora DSR proposal [1].

2.2 Interpolation

Interpolation approximates missing vectors as points on a curve fitted to those vectors surrounding the burst of loss. To some extent the order of the curve governs the accuracy of estimation. However, more complex forms of curves require additional information, such as derivatives, which can result in poor fitting to the data. In *linear interpolation* a straight line is fitted between the two vectors immediately surrounding the burst of loss, hence $N_{before}=1$ and $N_{after}=1$. The n^{th} vector of the burst is given by

$$\hat{\mathbf{x}}_n = \mathbf{x}_{\text{before}} + \frac{n}{\beta+1}\left(\mathbf{x}_{\text{after}} - \mathbf{x}_{\text{before}}\right) \qquad 1 \le n \le \beta \qquad (3)$$

Linear interpolation causes the velocity of the signal to become constant for the duration of the estimation which results in its acceleration becoming zero. Linear interpolation may also result in a discontinuity at the edges of the burst. A better approximation can often be made by fitting a non-linear segment between the points $\mathbf{x}_{\text{before}}$ and $\mathbf{x}_{\text{after}}$. This curve can be made continuous at the edges of the burst by matching the velocity of the curve at these points to that of the signal. This work has considered a number of methods for non-linear interpolation and has found that cubic *Hermite polynomials* give best estimates. The non-linear interpolation function for estimating the n^{th} lost vector in a burst of length β is,

$$\hat{\mathbf{x}}_n = \mathbf{a}_0 + \left(\frac{n}{\beta+1}\right)\mathbf{a}_1 + \left(\frac{n}{\beta+1}\right)^2 \mathbf{a}_2 + \left(\frac{n}{\beta+1}\right)^3 \mathbf{a}_3 \qquad 1 \le n \le \beta \qquad (4)$$

The multivariate coefficients, $\{\mathbf{a}_0,.., \mathbf{a}_3\}$, need to be calculated so that vectors at the start and end of the loss follow a smooth trajectory with the first derivatives of the polynomial being continuous at the start and end of the loss [6]. These coefficients can be computed from the two vectors preceding and following the burst of loss, $\mathbf{x}_{\text{before}}$ and $\mathbf{x}_{\text{after}}$, and their first derivates, $\mathbf{x}'_{\text{before}}$ and $\mathbf{x}'_{\text{after}}$. Expressing the interpolation function in terms of Hermite basis functions gives the estimate of the n^{th} feature vector within the burst as

$$\begin{aligned} \hat{\mathbf{x}}_n = {} & \mathbf{x}_{before}\left(1 - 3t^2 + 2t^3\right) + \mathbf{x}_{after}\left(3t^2 - 2t^3\right) \\ & + \mathbf{x}'_{before}\left(t - 2t^2 + t^3\right) + \mathbf{x}'_{after}\left(t^3 - t^2\right) \end{aligned} \qquad 1 \le n \le \beta \qquad (5)$$

where $t=n/(\beta+1)$, and derivates are approximated by $\mathbf{x}'_{before} = \beta\left(\mathbf{x}_{before} - \mathbf{x}_{before-1}\right)$ and $\mathbf{x}'_{after} = \beta\left(\mathbf{x}_{after+1} - \mathbf{x}_{after}\right)$. As estimation of derivatives requires two vectors either side of the burst, $N_{before}=2$ and $N_{after}=2$. If two vectors are not available the derivative is set to zero. In practice it was found that rapid fluctuations of the feature vector stream resulted in large estimates of the derivative components causing the interpolation to overshoot. Improved performance was achieved by reducing large derivative estimates by applying the following logarithmic compression to the vector differences,

$$f(x) = \text{sgn}(x)\log\left(|x| + 1\right) \qquad (6)$$

which gives $\mathbf{x}'_{before}$ and $\mathbf{x}'_{after}$ as,

$$\mathbf{x}'_{before} = \beta \times f\left(\mathbf{x}_{before} - \mathbf{x}_{before-1}\right) \qquad (7)$$

$$\mathbf{x}'_{after} = \beta \times f\left(\mathbf{x}_{after+1} - \mathbf{x}_{after}\right) \qquad (8)$$

3. INTERLEAVING

Interleaving is applied on the terminal device and serves to permute the order in which feature vectors are packetised such that bursts of loss are distributed amongst many shorter bursts. Formally, for a sequence of feature vectors, $\boldsymbol{X}$, where,

$$\boldsymbol{X} = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\text{N-1}}\} \qquad (9)$$

interleaving can be expressed as a permutation producing a re-ordered sequence, $\boldsymbol{X}'$, given as,

$$\boldsymbol{X}' = \{\mathbf{x}_{\pi(0)}, \mathbf{x}_{\pi(1)}, \mathbf{x}_{\pi(2)}, \ldots, \mathbf{x}_{\pi(\text{N-1})}\} \qquad (10)$$

where, $\pi(i)$, is interleaving function and gives the index of the vector to be output at the i^{th} time instance. The re-ordering made by the interleaving function requires feature vectors to be buffered prior to transmission which causes a delay in the end-to-end transmission time. The interleaving delay, δ, is defined as the maximum delay that any vector experiences before transmission,

$$\delta = \max_i\left(\pi^{-1}(i) - i\right) \qquad (11)$$

where π^{-1} is the inverse function of π. The ability of an interleaver to disperse bursts of loss is related to its spread. An interleaver has spread s if all pairs of vectors that are within s vector indexes of each other in the input sequence are separated by at least s vector indexes in the output sequence,

$$|x - y| \ge s \ \textit{ whenever } \ |\pi(x) - \pi(y)| < s \qquad (12)$$

A burst of packet loss of length β will be totally distributed (i.e. no concurrent packets will be lost) by an interleaver with spread s if $s \ge \beta$. For the case $s < \beta$ the interleaver will not be able to fully distribute the burst which will result in some consecutive packets being lost. The remainder of this section considers three forms of interleaver for application to DSR.

3.1 Optimal spread block interleavers

A block interleaver of size N operates by permuting the transmission order of a block of N feature vectors. This same permutation is applied to all subsequent blocks. Two block interleavers, π_{block1} and π_{block2}, [7] are considered optimal in terms of maximising their spread for given size and are given as,

$$\pi_{block1}(id + j) = (d - 1 - j)d + i \quad \text{where } 0 \le i,j \le d-1 \quad (13)$$

$$\pi_{block2}(id + j) = jd + (d - 1 - i) \quad \text{where } 0 \le i,j \le d-1 \quad (14)$$

where $d = \sqrt{N}$. It is interesting to observe that π_1 and π_2 form an invertible pair as $\pi_1 = \pi_2^{-1}$ and $\pi_2 = \pi_1^{-1}$. The delay and spread of these two interleavers is related to the square root of their size. From equations 13 (or 14), 11 and 12 the block interleaver delay, δ_{block}, and spread, s_{block}, are given as,

$$\delta_{block} = d^2 - d \quad \text{and} \quad s_{block} = d \quad (15)$$

3.2 Decorrelated block interleavers

The previous interleaver disperses burst-like packet loss by maximising spread according to equation 12. An alternative view of interleaving is that it is the process of decorrelating the order in which vectors are output in relation to their input order. In this view maximising decorrelation will minimise the resulting average burst lengths. A decorrelated block interleaver of size N consists of a permutation sequence of length N, defined by,

$$\pi_{decorrelated} = \{\pi(0), \pi(1), \cdots, \pi(N-1)\} \quad (16)$$

where the sequence $\pi(0...N-1)$ aims to maximise the *decorrelation measurement*, D_π,

$$D_\pi = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{|\pi(i) - \pi(j)|}{|i - j|} \quad (17)$$

The ability of an interleaver to distribute bursts of packet loss is related directly to its decorrelation value and is shown in an experiment where a set of 1000 block interleavers, each with random permutation sequences of length 16, is generated – $\{\pi_1$ to $\pi_{1000}\}$. A channel is simulated with packet loss rate α=50% and average burst length β=4 with each packet transporting 2 vectors. Figure 3a shows the output average burst length as a function of decorrelation value for each interleaver. Figure 3b shows the resulting speech recognition accuracy against decorrelation value.

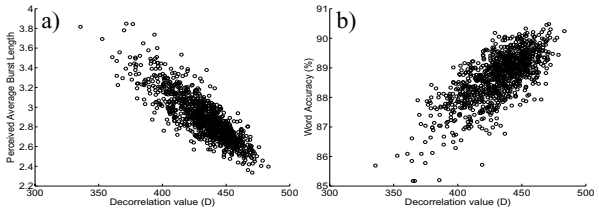

Figure 3: Decorrelation value against a) average burst length and b) word accuracy for 1000 random block interleavers.

The strong negative correlation in figure 3a shows that interleavers with high decorrelation values are more effective at distributing bursts of packet loss than those with lower decorrelation values. Figure 3b shows that interleavers with high decorrelation values enable higher recognition accuracy because of the shorter duration bursts over which estimation must operate.

For a block interleaver of size N, the selection of permutation sequence to maximise the decorrelation value is not elementary. The number of possible permutation sequences of length N is $N!$, hence a comprehensive state space search becomes impractical for higher degree interleavers. Heuristic search methods allow longer permutations to be created but do not guarantee that results will be optimal. The decorrelated interleavers used in this work have been selected using a greedy local search [8], where movement in the state space is defined by the swapping of two elements in the permutation sequence. Once a suitable sequence has been found its delay and spread can be determined from equations 11 and 12.

3.3 Convolutional interleavers

Convolutional interleavers can be modelled as an arrangement of shift registers each holding one feature vector [7]. Sequential input feature vectors are divided amongst different sub-sequences. Each sub-sequence consists of a different number of connected shift registers and hence imposes a different delay to the feature vectors stored in it. A convolutional interleaver of size N has $d = \sqrt{N}$ sub-sequences and takes the form,

$$\pi_{conv}(i) = i - d(i \mod d) \quad (18)$$

The delay, δ_{conv}, and spread, s_{conv}, of a convolutional interleaver are related to d and from equations 18, 11 and 12 are given as,

$$\delta_{conv} = d^2 - d \quad \text{and} \quad s_{conv} = d - 1 \quad (19)$$

4. EXPERIMENTAL RESULTS

The experimental results examine the effect that the different types of vector estimation and interleaving have on recognition accuracy for a variety of simulated channels. The recognition task for these experiments is the Aurora connected digit database [1]. Digits are modelled using 16-state, 3-mode HMMs, trained from a set of 8440 digit strings, using static MFCCs together with velocity and acceleration derivatives. The test set comprises 4004 noise-free digits strings (13,159 digits in total) which gives baseline accuracy of 98.5% with 95% confidence error bands of +/- 0.76% at 75% accuracy and +/- 0.38% at 95% accuracy. As per the ETSI standard, two vectors are carried by each packet.

Four channels were simulated by a 3-state Markov chain [4] to give a mixture of network conditions in terms of the packet loss rate, α, and average burst length, β. The channel parameters are shown in table 1 and include both high and low loss rates and long and short average burst lengths.

Channel	Loss rate, α	Av. Burst length, β	Baseline accuracy (no compensation)
A	10%	4	91.19%
B	10%	20	89.43%
C	50%	4	49.56%
D	50%	20	49.61%

Table 1: Simulated channel conditions.

4.1 Missing vector estimation

Experimental results, shown in table 2, measure the effect on recognition accuracy of the various estimation schemes described in section 2. At this stage no interleaving has been applied.

Method	*A*	*B*	*C*	*D*
No compensation	91.2	89.4	49.6	49.6
Repetition	94.7	90.2	76.0	53.8
NN Repetition	96.6	91.6	84.0	58.8
Linear interpolation	96.4	91.2	81.5	56.8
Hermite interpolation	96.4	90.7	80.4	54.7
Log Hermite interpolation	96.7	91.7	87.2	60.5

Table 2: Recognition performance for vector estimation schemes.

The results show that Hermite interpolation, with logarithmic compression of the first derivative, gives superior performance in all the channels tested. However, it should be noted that nearest neighbour repetition has similar performance and is the result of a less complex process. Hermite interpolation, without logarithmic compression, gives poor performance and demonstrates the importance of preventing large overshoots in estimation.

4.2 Interleaving

Based on the superior performance of log Hermite interpolation, figure 4 shows the effect of combing this with the three interleavers described in section 3 on the four channel conditions. For each class of interleaver the size is varied between 1 and 64 vectors. The interleaving size of 1 is equivalent to no interleaving and corresponds to the accuracy given in table 2. Equations 15 and 19 show that the delay of an interleaver and its spread are both functions of interleaver size. This means there is an inherent trade-off between word accuracy and delay; therefore results are presented as word accuracy against delay for each interleaver.

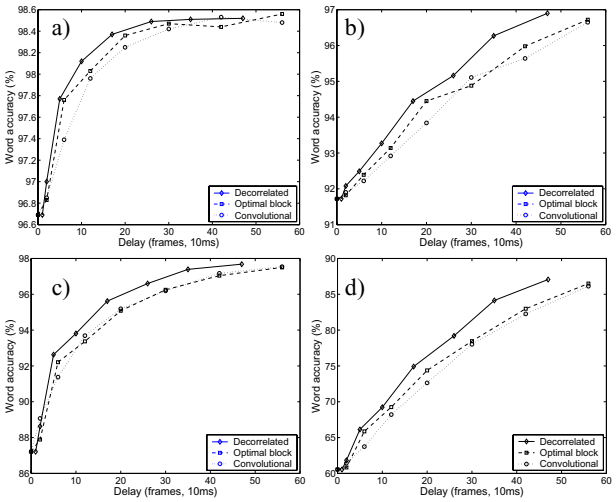

Figure 4: Recognition accuracy against delay as a function of size for various interleavers and channel conditions.

The figures show that interleaving feature vectors prior to transmission results in a significant increase in word accuracy, the magnitude of which is related to interleaver size. Figures 4a and 4c correspond to relatively short burst lengths and show the increase in accuracy levelling out as the size (and hence the spread) of the interleavers becomes sufficient to fully distribute the bursts. Increasing the interleaver size beyond this point gives no further increase in recognition accuracy. For longer burst lengths, shown in figures 4b and 4d, the interleaver size is not sufficient to fully distribute the bursts but does offer gains in accuracy. To restore performance for these longer burst lengths the interleaving delay would need to be considerably longer than shown. The figures also show that whilst all interleavers offer useful performance gains, the decorrelated interleaver generally results in slightly higher accuracy whilst imposing a smaller delay than the other interleavers.

5. CONCLUSIONS

This work has shown that packet loss can have a severe effect on recognition accuracy. Improvements can be made by replacing lost vectors with estimates based on received vectors surrounding the burst of loss. All five methods of estimation lead to improvements in recognition accuracy with nearest neighbour repetition and Hermite interpolation, with logarithmic compression, giving best performance. Testing on different channel conditions showed that estimation methods are able to recover performance even on very lossy channels provided burst lengths are reasonably short. At longer burst lengths estimation techniques become less effective due to the non-stationarity of the vector stream. To reduce burst lengths, and hence improve vector estimation, three types of interleaver have been considered. Experiments showed that increasing the size of the interleaver gave substantial increases in recognition performance, but at the expense of an exponential increase in delay. Analysis into the usability of speech recognition systems suggests that this delay should be less than 500ms [5]. Of the three interleavers tested, the decorrelated interleaver gave slightly superior recognition accuracy whilst imposing a lower delay. These results show that recognition accuracy is more affected by the average burst length of packet loss rather than the overall percentage of loss. This suggests that for robust recognition performance it is more important to distribute bursts of loss through techniques such as interleaving rather than attempting to reduce the overall percentage of lost packets.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) in this work.

7. REFERENCES

[1] ESTI document - ES 201 108 – STQ: DSR – Front-end feature extraction algorithm; compression algorithm, 2000

[2] Kim, H.K. and Cox, R.V., “A bitstream-based front-end for wireless speech recognition on IS-136 communication system”, IEEE Trans. SAP, Vol. 9, No. 5, pp. 558-568, 2001

[3] Milner B.P., James A.B. “Analysis and compensation of packet loss in distributed speech recognition using interleaving”, Proc. Eurospeech, 2003

[4] Milner, B.P., “Robust speech recognition in burst-like packet loss”, Proc. ICASSP, 2001.

[5] Boulis, C., et al, “Graceful degradation of speech recognition performance over packet-erasure networks”, IEEE Trans. SAP, vol. 10, No. 8, pp. 580-590, November, 2002.

[6] Vaseghi, S.V., “Advanced digital signal processing and noise reduction”, John-Wiley, second edition, 2000.

[7] Andrews K, Heegard C, Kozen D. “A theory of interleavers”, Technical report 97-1634, Computer Science Department, Cornell University, June 1997.

[8] Russell S, Norvig P., “Artificial Intelligence: A modern approach”, Prentice Hall, second edition, 2003.